CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis

Year: 2022 | *Citations: 1312* | *Authors: Erik Nijkamp, Bo Pang, Hiroaki Hayashi*

Abstract

Program synthesis strives to generate a computer program as a solution to a given problem specification, expressed with input-output examples or natural language descriptions. The prevalence of large language models advances the state-of-the-art for program synthesis, though limited training resources and data impede open access to such models. To democratize this, we train and release a family of large language models up to 16.1B parameters, called CODEGEN, on natural language and programming language data, and open source the training library JAXFORMER. We show the utility of the trained model by demonstrating that it is competitive with the previous state-of-the-art on zero-shot Python code generation on HumanEval. We further investigate the multi-step paradigm for program synthesis, where a single program is factorized into multiple prompts specifying subproblems. To this end, we construct an open benchmark, Multi-Turn Programming Benchmark (MTPB), consisting of 115 diverse problem sets that are factorized into multi-turn prompts. Our analysis on MTPB shows that the same intent provided to CODEGEN in multi-turn fashion significantly improves program synthesis over that provided as a single turn. We make the training library JAXFORMER and model checkpoints available as open source contribution: https://github.com/salesforce/CodeGen.